Variable Selection for Gaussian Process Models in Computer Experiments

Crystal Linkletter[1], Derek Bingham[1], Nick Hengartner[2], Dave Higdon[2], and Kenny Ye[3]

[1]Department of Statistics and Actuarial Science
Simon Fraser University

[2]Statistical Sciences Group
Los Alamos National Labs

[3]Department of Applied Math and Statistics
SUNY, Stony Brook

Abstract

In many situations, simulation of complex phenomena requires a large number of inputs and is computationally expensive. Identifying the inputs which most impact the system so that these factors can be further investigated can be a critical step in the scientific endeavor. In computer experiments, it is common to use a Gaussian spatial process to model the output of the simulator. In this article, we introduce a new, simple method for identifying active factors in computer screening experiments. The approach is Bayesian and only requires the generation of a new inert variable in the analysis; however, in the spirit of frequentist hypothesis testing, the posterior distribution of the inert factor is used as a reference distribution against which the importance of the experimental factors can be assessed. The methodology is demonstrated on an application in material science, a computer experiment from the literature, and simulated examples.

KEY WORDS: Computer simulation; Latin hypercube; Random field; Screening; Spatial Process.

1 Introduction

Rapid growth in computer power has made it possible to study complex physical phenomena that might otherwise be too time consuming or expensive to observe. Scientists are able to adjust inputs to computer simulators (or computer codes) in order to help understand their impact on a system. Many such computer simulators require the specification of a large number of input settings and are computationally demanding. As a result, only a limited number of simulation runs tend to be carried out. Scientists must therefore select the simulation trials judiciously and perform a designed computer experiment (or simply a computer experiment).

One main goal of experimentation, particularly in its early stages, is to determine the relative importance of each input variable in order to identify which have a significant impact on the process being studied. Since there can be many inputs into a computer code, an important problem is the identification of the most active factors. This is often referred to as *screening* (e.g., Wu and Hamada, 2000) and is the main focus of this article.

Most computer experiments are unique in that the response has no random error component. That is, replicates of the same inputs to the computer code will yield the same response. To deal with this, Sacks et al. (1989a, 1989b) proposed modelling the response from a computer experiment as a realization from a stochastic process. This allows for estimates of uncertainty in a deterministic computer simulation. Welch et al. (1992) also used this model in their approach to screening in complex computer experiments. We focus specifically on modelling the computer simulated response as a Gaussian process with a spatial correlation structure. This model is particularly attractive since it fits a very large class of response surfaces. In addition, a white-noise component is introduced that allows for random error such as small errors in computational convergence in the computer model. The flexibility of this model does not come without a price, however. Choosing which predictor variables are active in the spatial model presents a challenging variable selection problem.

In this article, a new approach for screening when using a spatial process model is presented. The method, reference distribution variable selection (RDVS), is used to identify a subset of factors to be examined more closely in later stages of experimentation. To carry out RDVS, the experimental variables are augmented with a factor which is known to be inert, and a Bayesian analysis is repeated several times. Rather than basing screening decisions solely on the posterior distributions of the relevant model parameters, the relevant posterior distributions are compared to a reference distribution derived from the inert factor inserted into the analysis. Parameters with more extreme posterior distributions relative to the reference distribution are deemed important. Though the model fitting is fully Bayesian, we use a reference distribution to identify important factors, giving the variable selection approach a frequentist flavor.

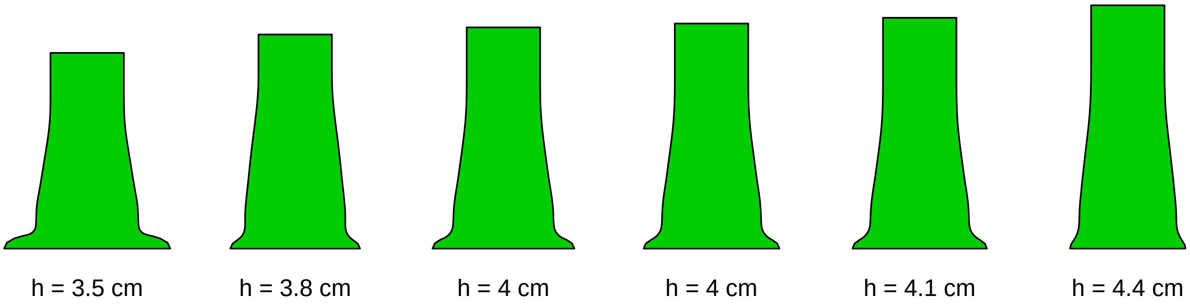
The RDVS methodology has many features that make it attractive for screening. First, implementation of the methodology requires only the inclusion of one extra variable, which contributes little to the complexity of the analysis. Second, the prior distribution placed on the inert factor is the same as the prior placed on the real process variables, making them comparable with no additional tuning required. In essence, the added factor self-calibrates itself to the other inert factors, making the analysis relatively robust to prior specification.

A brief outline of the paper is as follows. In the next section, we consider screening for a cylinder deformation computer simulator. Next, the Gaussian process model used in the analyses of computer experiments is presented in Section 2. In Section 4, the RDVS methodology is introduced in detail and a variable selection criterion is proposed. The performance of the methodology is demonstrated in Section 5 by applying it to a problem from the literature (Schonlau and Welch, 2005) and simulated examples, followed by an exploration of its robustness to prior choice in Section 6. Finally, in Section 7, we use RDVS to successfully identify a subset of important factors in a real application. We finish the paper with some concluding remarks in Section 8.

2 Cylinder Deformation Application

Detailed computer simulation of physical processes plays an important role in the development and understanding of physics-based mathematical models. One of the applications we have been working with at Los Alamos National Laboratory (LANL) is a finite element code that simulates a high velocity impact of a cylinder (hereafter referred to as the Taylor cylinder experiment). In this experiment, a copper cylinder (length = 5.08 cm, radius = 1 cm) is fired into a fixed barrier at a velocity of 177m/s. The resulting impact deforms the cylinder according to the plastic deformation model of Preston, Tonks, and Wallace (2003), the PTW model. This model is governed by 14 parameters (factors), which in essence specify how much energy is required to crush each cylinder. Figure 1 shows a sample of cylinder deformations corresponding to a range of settings for these input parameters.

Figure 1: Collection of simulated cylinders ranging from most compressed to the least taken from the set of 118 simulations of the Taylor cylinder test.



The PTW model was developed to be applicable for a wide range of strain rates, and, in general, all of the factors play an important role in simulating the the deformation. Indeed, this is why they were included as inputs to the computer code. However, a computer experiment frequently exercises the simulator over a limited range of physical conditions (e.g., velocities or strain rates). Over this range, the simulator response is often dominated by a very limited number of input parameters.

At the input velocity of 177m/s used for this experiment, it is expected that deformation will only be affected by a subset of the 14 input parameters. Furthermore, the Taylor cylinder experiment is only a small component of broader experimentation, so reducing the number of factors to carry on to further experiments is beneficial. The goal of this study is to identify which factors most significantly impact the deformation (i.e., screening) over the reduced input space of the complex computer simulator. In the following sections, we introduce the most common approach for modelling the response from a computer code and propose a new approach for variable selection. We re-visit this example in Section 7.

3 Gaussian Process Model

To model the response from a computer experiment, we use a Bayesian version of the Gaussian process (GP), first proposed by in this setting Sacks et al. (1989a). Our formulation is tailored to the types of simulation models we have been working with at the LANL, such as the Taylor cylinder experiment described in the previous section. In many cases, the computer codes simulate well understood physical processes, taking in a number of input parameters and frequently producing a highly multivariate output. For the purpose of this paper, focus will be on a univariate summary of interest from the output. Simulator inputs typically describe initial conditions as well as physical parameters such as material strength or equations of state. In this article, we make no distinction between parameter type, although this distinction is important in model calibration (e.g. Kennedy and O'Hagan, 2001; Goldstein and Rougier, 2004).

Although a simulator generally requires a large number of inputs that play an important role in emulating a physical process over a broad range of conditions, computer experiments often only exercise the simulator over a limited range, where only a few factors dominate the response. It is worth noting that the specification of the input setting range can greatly influence whether or not a particular input is selected as active. In the applications we have considered, there is typically experimental data or specific application requirements of the simulator that lead to specifying this range, so these input ranges can be taken as fixed and appropriate. The goal of the screening experiment is to identify these active factors – or inputs – in light of factor sparsity (Box and Meyer, 1986). The GP provides a flexible framework for response surface modelling, but this flexibility makes deciding which factors are active and which are inert more challenging.

The input to the computer code, X, is an $n \times p$ matrix built by stacking n input vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, each of length p. The corresponding outputs for the n simulation runs are held in the n-dimensional vector $y(X)$. As mentioned, this is often a one-dimensional summary of a multivariate output. To facilitate prior specification, the response is standardized so that it has a mean of zero and a variance of one. The design space is also transformed so that the input settings range over the p-dimensional unit cube $[0, 1]^p$.

We model the (centered and scaled) simulator response, $y(X)$, as the sum of a GP, $z(X)$, which depends on the design matrix, X, and independent white noise. That is,

$$y(X) = z(X) + \epsilon, \tag{1}$$

where ϵ is a mean zero white noise process with variance $1/\lambda_\epsilon$, independent of $z(X)$. The GP,

$z(X)$, is specified to have mean zero and covariance function

$$\operatorname{Cov}(z(\mathbf{x}_i), z(\mathbf{x}_j)) = \frac{1}{\lambda_z} \prod_{k=1}^{p} \rho_k^{2^{\alpha_k} |x_{ik} - x_{jk}|^{\alpha_k}}. \tag{2}$$

Here, x_{ik} denotes the k^{th} input value for the i^{th} simulation trial.

There are a few features about the correlation function in (2) that one should notice. Under the re-parameterization $\rho_k = e^{-(1/2)^{\alpha_k}\theta_k}$ $(\theta_k > 0)$, this correlation function is the same as that suggested by Sacks et al. (1989a). We prefer the parameterization in (2) because it facilitates posterior exploration via Markov chain Monte Carlo (MCMC). In addition, interpretation is straightforward: if ρ_k is large (i.e. close to one), the process does not depend on factor k. Therefore, estimation of the ρ_k's helps to determine which of the input variables, or factors, may be active for the given investigation. In general, specifying $\alpha_k = 2$ in (2) is a reasonable simplifying assumption for our simulator-based applications. This is because the simulator response to input changes is nearly always smooth in our experience. Also, any roughness in response that would suggest taking $\alpha < 2$ is typically from numerical "jitter" and is better accounted for in the error process ϵ. The stipulated covariance given in (2) is separable in the sense that it is the product of component-wise covariances. This enables one to handle a large number of inputs since each input dimension k requires only one additional parameter, ρ_k. In addition, this simplified covariance structure still accommodates multi-way interactions in $z(X)$.

The GP, $z(X)$, accounts for most of the variation in the simulation output, while the error term ϵ is meant to account for local roughness in the simulator response, typically due to numerical effects such as gridding, tabular function representation, or convergence tolerance. Because the observed simulation output is standardized to have mean zero and variance one, we wish to specify a prior for λ_z that encourages its value to be close to one. One way to do this is to take λ_z to have a $\Gamma(a_z = 5, b_z = 5)$ prior (i.e. $\pi(\lambda_z) \propto \lambda_z^{a_z-1} e^{-b_z \lambda_z}$). In our experience, if $z(X)$ is adequately approximating the simulator response, the error term should account for only a small amount of the response standard deviation. Thus, the prior specification distribution for λ_ϵ should have a relatively large mean with a constraint to prevent λ_ϵ from being too small. One reasonable choice is a gamma prior, $\Gamma(a_\epsilon = 2.5, b_\epsilon = 0.025)$, with the condition that $\lambda_\epsilon > 5$, so that $\pi(\lambda_\epsilon) \propto \lambda_\epsilon^{a_\epsilon - 1} e^{-b_\epsilon \lambda_\epsilon} I[\lambda_\epsilon > 5]$. Under this prior, the error term is expected to account for only about 10% of the response standard deviation, and with the given constraint, never more than about 45%.

The prior specification for $\rho = (\rho_1, \ldots, \rho_p)^T$ is motivated by the variable selection priors from the regression context (e.g., George and McCulloch, 1993; Clyde, 1999). Each component of ρ is given an independent mixture prior of a standard uniform and a point mass at one:

$$\pi(\rho_k) = \gamma I[0 \leq \rho_k \leq 1] + (1-\gamma)\delta_1(\rho_k). \tag{3}$$

Here, γ is the prior probability that input k is active and $\delta_1(\cdot)$ denotes a point mass at one. The model prior for ρ is the product of the priors on the independent components. This specification is particularly attractive in the screening context because the mixture probability can be chosen to reflect prior beliefs on the number of active factors, thereby incorporating the effect sparsity assumption into the prior. In the examples explored later, we specify $\gamma = 1/4$ to encode a prior belief that about one quarter of the p inputs will be active (e.g., factor sparsity). By specifying a value of γ that does not depend on the number of inputs, prior beliefs are not impacted by the addition of the inert variable for RDVS.

The likelihood implied by (1) for the response given $z(X)$ and λ_ϵ is

$$L(y|z, \lambda_\epsilon) \propto \lambda_\epsilon^{\frac{n}{2}} \exp\{-\tfrac{1}{2}\lambda_\epsilon (y-z)^T(y-z)\}.$$

Using a Gaussian process model for the response surface $z(X)$ yields the prior

$$\pi(z|\lambda_z, \rho) \propto \lambda_z^{\frac{n}{2}} |R(\rho)|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}\lambda_z z^T R(\rho)^{-1} z\},$$

where $R(\rho)$ is an $n \times n$ matrix with entries

$$R_{ij}(\rho) = \prod_{k=1}^{p} \rho_k^{4(x_{ik}-x_{jk})^2}.$$

Recall that this is the correlation function given in (2) with $\alpha_k = 2$.

This likelihood, along with the prior distributions for $z, \lambda_\epsilon, \lambda_z$ and ρ as given, leads to the posterior density

$$\pi(z, \lambda_\epsilon, \lambda_z, \rho|y) = L(y|z, \lambda_\epsilon) \times \pi(z|\lambda_z, \rho) \times \pi(\lambda_\epsilon) \times \pi(\lambda_z) \times \pi(\rho).$$

After integrating out z, the posterior density for λ_ϵ, λ_z and ρ is found to be

$$\begin{aligned}\pi(\lambda_\epsilon, \lambda_z, \rho|y) \;\; \propto \;\; & \left|\frac{1}{\lambda_\epsilon}I_n + \frac{1}{\lambda_z}R(\rho)\right|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}y^T(\frac{1}{\lambda_\epsilon}I_n + \frac{1}{\lambda_z}R(\rho))^{-1}y\right\} \times \lambda_\epsilon^{a_\epsilon - 1}e^{-b_\epsilon\lambda_\epsilon} \\ & \times\lambda_z^{a_z-1}e^{-b_z\lambda_z} \times \pi(\rho), \end{aligned} \tag{4}$$

in which I_n denotes the $n \times n$ identity matrix. Realizations from this $(p + 2)$-dimensional posterior distribution can be drawn using a standard MCMC algorithm, which only requires Metropolis updates for implementation. In particular, it is the realizations of $\rho_k, \; k = 1, \ldots, p$ that are used for making screening decisions. Ideally, one can find a cut-off value for the ρ_k's which can then be used to decide if a factor is active or inert in the spirit of a frequentist hypothesis test. In fact, in the sequel, we do just that.

4 RDVS for Spatial Models

As mentioned, the GP provides a flexible framework for response surface modelling. It produces a much broader class of potential response surfaces than the classical linear or polynomial regression models, and easily adapts to the presence of non-linearity and interactions. However, this flexibility makes deciding which factors are active and which are inert challenging. When there are p factors, there are 2^p possible combinations of factors. A good discussion on assigning model priors is given by Chipman, George and McCulloch (2001). A fully Bayesian implementation for the screening of active factors often requires one to stipulate a prior on all 2^p possible sub-sets, which is not always straightforward. In addition, variable selection decisions in this context are often subjective and sensitive to prior model specification.

In this section, a new simple method for assessing the significance of factors in a GP is introduced. It is worth noting that although RDVS is introduced with reference to the GP, we believe its application is much broader and can be adapted to other models. Our approach to identifying which individual estimates of ρ_k are small enough (far enough from one) to be deemed as evidence of a significant factor parallels a frequentist's approach to model selection. The central issue is to identify a reference distribution and selection criterion that can be used to assess the importance of the experimental factors.

To outline the approach, consider the spatial model in (1). Since it is unknown which experimental factors are important (indeed, finding them is the goal of the analysis), gauging the relative magnitudes of the ρ_k's can be difficult. RDVS serves as a remedy to this problem. To implement, an additional variable which is known to be inert – and thus has no impact on the response – is appended to the design matrix. Now the experimenter knows how an inert factor behaves, and therefore has a benchmark against which the experimental factors can be compared. We propose to use the posterior distribution of the inert, or null, variable as a reference distribution to decide which of the real factors are important.

Consider the $n \times p$ design matrix X, where the i^{th} row of the design matrix represents the level settings of the p continuous covariates for the i^{th} trial. An augmented $n \times (p+1)$ design matrix is constructed by adjoining to X one additional column, $X^{\dagger} = (x_{1(p+1)}, x_{2(p+1)}, \ldots, x_{n(p+1)})^T$. To mimic the p real covariates, the elements of the additional column vector, $X^{\dagger}$, range from 0 to 1 (recall that the original design matrix X has been scaled to this range). Ideally, the column vector $X^{\dagger}$ is orthogonal to each set of columns in X. In practice, this is unlikely to be the case, and so a way to select $X^{\dagger}$ is discussed shortly.

We emphasize that, by construction, the augmented variable is not a true experimental variable and has no impact on the response. The analysis proceeds as if there are $p+1$ factors, but in this case it is known that the added factor is inert. Therefore, the posterior distribution of ρ_{p+1} is the posterior distribution of the correlation parameter for an inert variable. Since the variable selection problem amounts to deciding which variables have an impact on the response that is distinguishable from noise, the posterior distributions of the experiment variables can be compared to that of the added variable to decide which variables can be claimed as active. That is, similar to frequentist hypothesis testing, the posterior distribution of the added variable is used as a reference distribution to assess the importance of the ρ_k's corresponding to the experimental factors. The key feature of this approach is that it makes judging the actual size or ranks of the ρ_k's unnecessary (i.e. the experimenter does not need to specify an arbitrary value below which ρ_k is considered to be sufficiently less than one). This is beneficial since which values of ρ_k are "small" is dependent on the particular data at hand. The only judgement that is necessary for RDVS is whether or not the posterior distributions of the experimental factors are distinguishable from the posterior distribution of the inert variable.

As noted, the ideal choice of $X^\dagger$ would be orthogonal to all sets of columns in X so that the posterior distribution of ρ_{p+1} would not be impacted by the choice of level settings of the other experiment factors. Such a column is unlikely to exist. To address this, we randomly sample $X^\dagger$ from the design space of X, and perform the above GP model analysis. Since $X^\dagger$ may be correlated with some columns in X, this procedure is repeated several times and the posterior distributions of the added inert variables from each iteration are combined to form one reference distribution corresponding to that of a null variable. This has the effect of averaging over all added columns.

There are a variety of ways one could imagine comparing the posterior distributions of the experiment factors to that of the null variable. One possible approach is to use the individual realizations at each step of the MCMC. For example, it is known that ρ_{p+1} should be close to one by construction, and ρ_k corresponding to an active variable should be less than one. Thus, the difference $\rho_k - \rho_{p+1}$ should be negative for most realizations of the MCMC for an active factor, while centered around zero for an inert factor. Thus, at the l^{th} step of the MCMC, one could compute $\rho_k - \rho_{p+1}$ for $k = 1, ..., p$, and after the MCMC has converged have p reference distributions of $\rho_k - \rho_{p+1}$. As a rule, a percentile of the reference distribution could be used as a cut-off for making variable selection decisions. That is, if a chosen percentile of the k^{th} reference distribution, $\rho_k - \rho_{p+1}$, is less than zero, then factor k is active. The difficulty with this approach is specifying and interpreting an appropriate cut-off for decision making.

Instead, consider the following approach. Each time the inert factor is added to the design matrix, X, and the analysis is performed, summarize the posterior distribution of ρ_{p+1} by its median, $\tilde{\rho}_{p+1}$. The process of inserting an inert variable, running the MCMC, and saving the posterior median of ρ_{p+1} is repeated many times. From this, an estimate of the distribution for the posterior median of a correlation parameter corresponding to an inert variable is obtained. In addition, every realization of ρ_k, $k = 1, \ldots, p$, is recorded at each step of the MCMC for the true experimental factors. The posterior median over all realizations for each of the ρ_k can be compared to the reference distribution of the inert factor median to assess the importance of factor k. The posterior estimates $\tilde{\rho}_k$ can also be used for prediction, etc., but that is beyond the scope of this article.

The disadvantage of this approach is the increase in computational time – though not the complexity – since the MCMC must be run many times in order to construct the reference distribution. However, the approach has many advantages. If $\tilde{\rho}_k$ is compared to, for example, the 5^{th} percentile of the null distribution for posterior medians, a frequentist's interpretation of importance can be used (i.e. one would expect to falsely identify an inert factor as significant approximately five percent of the time). For screening experiments, one typically would prefer to err on the conservative side, so we propose using the 10^{th} percentile of the null distribution as a cut-off. By using this approach, the posterior distributions of the ρ_k can be compared and assessed.

To summarize, RDVS entails the following steps:

1. Augment the experiment design, X, by creating a new design column corresponding to a variable with no significant effect. The level settings of the added inert factor are selected at random and cover the same design region as the original experimental factors.

2. Find the posterior median of ρ_{p+1}.

3. Repeat steps 1 and 2 m times. Obtain a distribution for the posterior median of a null effect to be used as a reference distribution.

4. Compare the posterior medians $\tilde{\rho}_k$ of the experimental variables to the reference distribution to assess their importance. The percentile of the reference distribution used for comparison reflects the rate of falsely identifying inert effects as active.

5 Simulated Examples

To illustrate the performance of RDVS, we have chosen three simulated examples of varying complexity. Using known functions allows us to evaluate the methodology. For all of the examples, the design matrix used is a 54-run Latin hypercube design with $p = 10$ input variables. Latin hypercube designs are a popular choice (such designs were introduced by McKay et al. (1979) specifically for computer experiments) because they can be generated with minimal computational effort and fill the design space relatively well. In addition, when the sample inputs of such a design are projected into any one dimension, complete stratification is achieved.

The particular design used in these examples has the additional property that the minimum pairwise distances in each two-dimensional projection is approximately maximized, yielding a space-filling design in each of the $p(p-1)/2$ two-dimensional projections of the design space. After the three simulated examples, we will look at a larger example that has been used to illustrate other screening techniques from the literature for comparison.

EXAMPLE 1:

The first example is meant to demonstrate the performance of the RDVS methodology for a simple case. To begin, data are generated from the linear model

$$y(\mathbf{x}_i) = 0.2x_{i1} + 0.2x_{i2} + 0.2x_{i3} + 0.2x_{i4} + e_i, \tag{5}$$

where $e_i \sim N(0, \sigma^2)$ with $\sigma = 0.05$. After generation, the response is standardized to have mean zero and standard deviation one. For the simulation study, data are generated from the linear model given in (5) 1000 times and the important factors using RDVS are identified at each iteration of the simulation. For this example and each of the subsequent examples, $m = 100$ is used in step 3 of the algorithm.

For illustration, consider in detail one iteration of the simulation study. First, a response is generated as described above. To implement RDVS, an inert variable (i.e. an 11^{th} factor) is added to the design, with levels randomly selected from the design region of the original ten experimental inputs. The GP previously described in Section 3 is used to model the response surface. As mentioned, for all examples γ in (3) is taken to be 1/4 (in general, γ should be chosen to reflect the user's prior beliefs on effect sparsity).

Using the augmented design matrix, 600 iterations of the MCMC algorithm are run to generate posterior realizations of the ρ_k, $k = 1, \ldots, 11$, under the GP model, with the first 100 discarded as burn-in. The augmentation procedure and MCMC implementation is repeated $m = 100$ times. We find this is sufficient to obtain a reasonable estimate of the distribution for the posterior median of ρ_{11}. All 50,000 realizations of ρ_k for the ten experiment inputs are saved, and the posterior median of the correlation parameter for the inert variable is obtained. The combined 100 posterior medians $\tilde{\rho}_{11}$ form the reference distribution to be used for variable selection.

Figure 2: Posterior distributions of ρ_k for one iteration of the simulation study in Example 1.

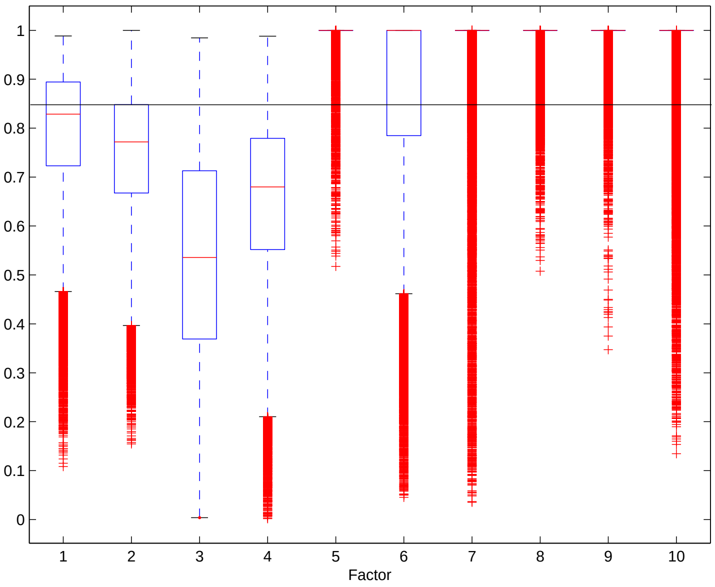


Figure 2 shows boxplots of the posterior realizations of ρ_k $(k = 1, \ldots, 10)$ obtained from the MCMC corresponding to one iteration of the simulation study. The 10^{th} percentile of the reference null posterior median distribution is indicated by the solid horizontal line on the figure. There are some features of Figure 2 worth noting. As usual, the boxes of the boxplots denote the first, second and third quartiles of a distribution. One can see in this plot that for this data, the posterior distribution of an inert factor, such as factor 5, is pushed up against one. Indeed, for this factor, the upper three quartiles of the posterior distribution are all one. The "tail" on the distribution shows the range of the small fraction of posterior realizations that are less than one. This pattern is also observed for the other inert factors to varying degrees. Conversely, the posterior distribution of an active factor (e.g. factor 1) is centered far less than one.

By just inspecting these boxplots, an experimenter would likely correctly identify the first four inputs as having a significant impact on the response because the posterior medians are all much less than 1 relative to the other factors. Looking at Figure 1, one may be tempted to also declare factor 6 active, however, the posterior median for this factor is exactly 1. If the more formal rule of comparing the posterior distributions of ρ_k for the experimental variables to the 10^{th} percentile of the null median distribution is followed, the first four inputs are indeed

Table 1: Proportion of times each factor is identified as important in 1000 generations of the linear function given in (5).

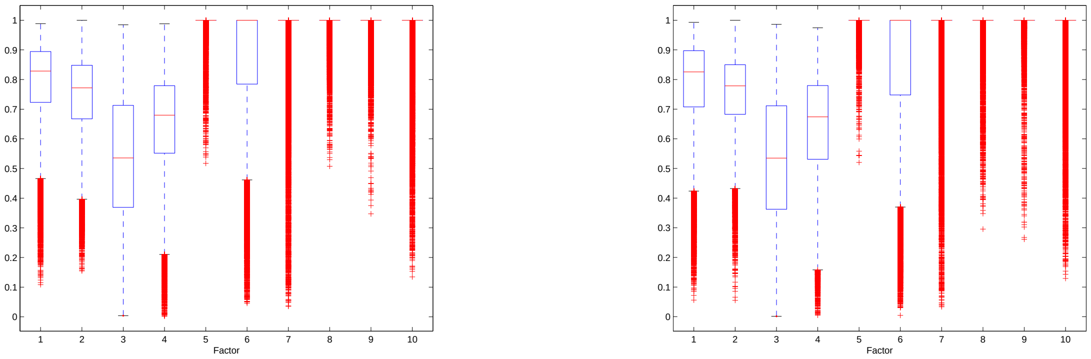
	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.619	0.618	0.717	0.631	0.030	0.034	0.021	0.074	0.051	0.051
10^{th}	0.852	0.855	0.910	0.880	0.061	0.064	0.053	0.137	0.076	0.102
15^{th}	0.947	0.954	0.973	0.955	0.079	0.091	0.080	0.173	0.108	0.135

correctly identified as being important. Thus, for this iteration of the simulation study, the decision is made to declare the first four inputs as active and the remaining factors as inert.

Table 1 summarizes the results for 1000 simulations. The performance of the approach is investigated using the 5^{th}, 10^{th}, and 15^{th} percentiles of the reference distribution as cut-off points. The results show that RDVS does well at correctly identifying the active factors in this simple example, as would be expected. It can also be seen from Table 1 that the false identification of inert inputs as active is at the expected level corresponding to the percentile used for decision making.

Before continuing, we make a brief digression back to the iteration of the simulation study explored in detail throughout this example. One might question if the addition of the extra variable for RDVS has an impact on the posterior distribution of the experimental variables. To explore this point, the MCMC analysis is repeated on this same response without adding the inert factor. Figure 3(a) shows the posterior distributions of the experimental variables when the extra factor is added, while Figure 3(b) shows the same distributions generated without using an augmented design matrix. The similarity of these plots suggests that there is no obvious altering of the experimental posterior distributions as a side effect of the methodology. Furthermore, inspection of the differences between posterior medians corresponding to the two approaches (with and without augmentation) showed an average difference of only 2.85×10^{-4}.

In order to explore the size of effects the RDVS selection method is able to detect, consider repeating this simulation study with a slightly more complex linear function. The response is now generated according to a linear function with decreasing coefficients on the first eight



(a) With augmentation

(b) Without augmentation

Figure 3: Posterior distributions of the experimental variables.

Table 2: Proportion of times each factor is identified as important in 1000 generations of the linear function given in (6).

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.679	0.180	0.062	0.025	0.016	0.023	0.017	0.031	0.009	0.036
10^{th}	0.889	0.379	0.133	0.058	0.034	0.051	0.035	0.067	0.030	0.094
15^{th}	0.959	0.540	0.217	0.092	0.061	0.098	0.065	0.107	0.063	0.149

inputs:

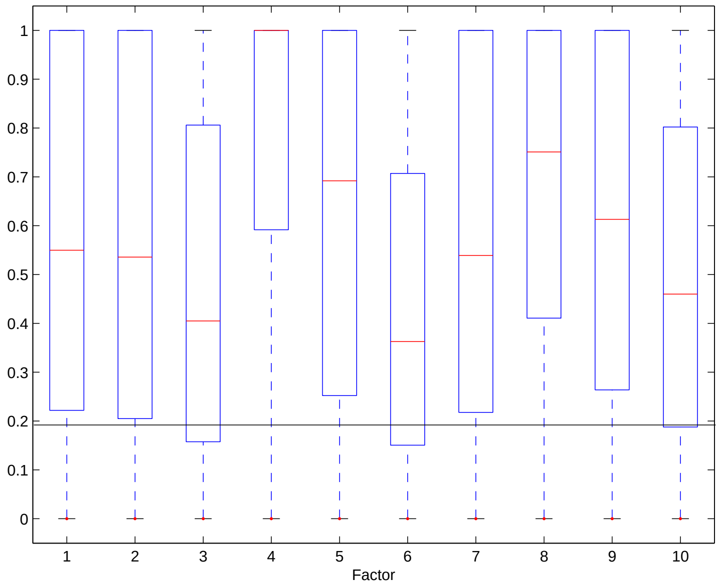
$$y(\mathbf{x}_i) = 0.2x_{i1} + \frac{0.2}{2}x_{i2} + \frac{0.2}{4}x_{i3} + \frac{0.2}{8}x_{i4} + \frac{0.2}{16}x_{i5} + \frac{0.2}{32}x_{i6} + \frac{0.2}{64}x_{i7} + \frac{0.2}{128}x_{i8} + e_i, \qquad (6)$$

where again $e_i \sim N(0, \sigma^2)$ with $\sigma = 0.05$. After generation, the response is standardized to have a mean zero and standard deviation of one. Table 2 gives the results for 1000 simulations of this response. From these results, it can be seen that the first factor is still easily identified as active, which is consistent with the previous results. In addition, the second and third factors are detected as active more often than would be expected by chance, while the remaining inputs (which have relatively small coefficients) are determined to be inert for the most part.

EXAMPLE 2:

For our second example, we explore how well RDVS can correctly identify a complete lack of signal. Welch et al. (1992) observed it is difficult to distinguish between a model with no

Figure 4: Posterior distributions of ρ_k for one iteration of the simulation study in Example 2.



active factors and one with all active factors. Indeed, the sequential likelihood approach to screening they proposed does not distinguish between these two models. In this case, because RDVS decisions are made by making comparisons with an inert variable, it is anticipated the methodology will be able to correctly detect a lack of activity amongst the experimental variables when none exists. For this example, the response is generated as random noise. That is, $y(\mathbf{x}_i) = e_i$, where $e_i \sim N(0, \sigma^2)$ with $\sigma = 0.05$, and analysis proceeds as in Example 1. Figure 4 shows boxplots corresponding to one iteration of this simulation study.

Note that in this plot it appears that all factors have correlations much less than one and seem to be significantly impacting the response. This is because the amount of variability that can be attributed to random noise is restricted in the model, and therefore the GP tries to interpolate a signal through most of the "jitter". In this case, based on a subjective examination of the boxplots, an experimenter would likely incorrectly declare all the ρ_k's to be less than one (and therefore important). When RDVS is used, however, the extra null factor added for the analysis looks and behaves like all the other inert factors, as indicated by the low value of the 10^{th} percentile of the reference distribution drawn in the figure. As a result, when the RDVS decision rule is used, the correct variable selection decisions are made. This illustrates the point that RDVS is based on comparisons between the experimental factors and the inert

Table 3: Proportion of times each factor is identified as important in 1000 generations of random noise.

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.003	0.013	0.004	0.007	0.006	0.008	0.001	0.006	0.003	0.005
10^{th}	0.012	0.039	0.009	0.013	0.016	0.022	0.010	0.017	0.011	0.013
15^{th}	0.033	0.064	0.032	0.039	0.029	0.041	0.023	0.027	0.027	0.031

factor, not on the actual values of the realized ρ_k's. The results from 1000 simulations are given in Table 3. It can be seen from these results that RDVS performs extremely well in this setting.

EXAMPLE 3:

For the third example, the data is generated according to

$$y(\mathbf{x}_i) = sine(x_{i1}) + sine(5x_{i2}) + e_i, \tag{7}$$

where again, $e_i \sim N(0, \sigma^2)$ with $\sigma = 0.05$ and the response is standardized. This function is slightly more complex than the others considered because factor 1 and factor 2 impact the response quite differently over their $[0, 1]$ ranges.

Figure 5 shows the posterior distribution of ρ_k, $k = 1, \ldots, 10$, for one iteration of this simulation. For this data, the posterior distributions corresponding to the inert variables are all pushed tightly against one. As it should, the added null variable mimics this behavior, as can be seen by looking at the 10^{th} percentile of the distribution for posterior medians of inert variables drawn on the figure. As a result, RDVS correctly detects that the distributions for ρ_1 and ρ_2 look discernibly different than the distribution for ρ of an inert factor. Table 4 summarizes the results for 1000 simulations. For this example, RDVS does very well at identifying factors 1 and 2 as having a significant impact on the response.

EXAMPLE 4:

We now illustrate the methodology a more challenging example for variable selection. Consider the "Wonderland" computer simulator (Milik, Prskawetz, Feichtinger, and Sanderson, 1996) based on a mathematical model for exploring strategies for sustainable global development.

Figure 5: Posterior distributions of ρ_k for one iteration of the simulation study in Example 3.

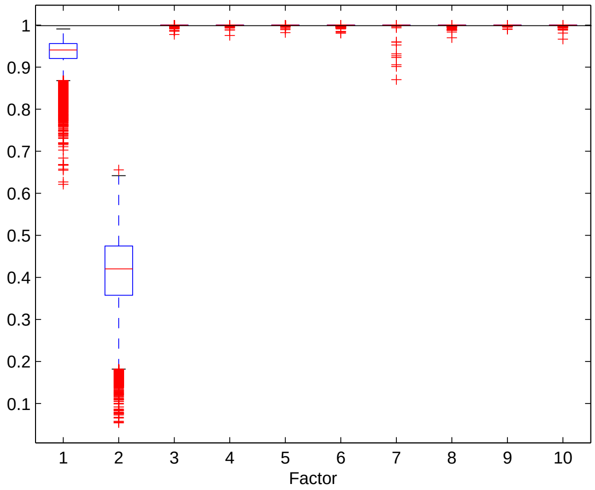


Table 4: Proportion of times each factor is identified as important in 1000 generations of the response given by (7).

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	1.000	1.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001
10^{th}	1.000	1.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001
15^{th}	1.000	1.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001

There are 41 inputs that generally fall into the three categories: population, economic, and environmental factors. Each combination of settings for the 41 factors represents a strategic policy for maintaining a sustainable environment. The output of the computer code is a human development index (HDI), where high values of the index indicate a more desirable, sustainable system. The analysis we consider here is based on a computer experiment consisting of 500 runs of a space-filling (e.g., see Johnson, Moore and Ylvisaker, 1990) Latin hypercube design.

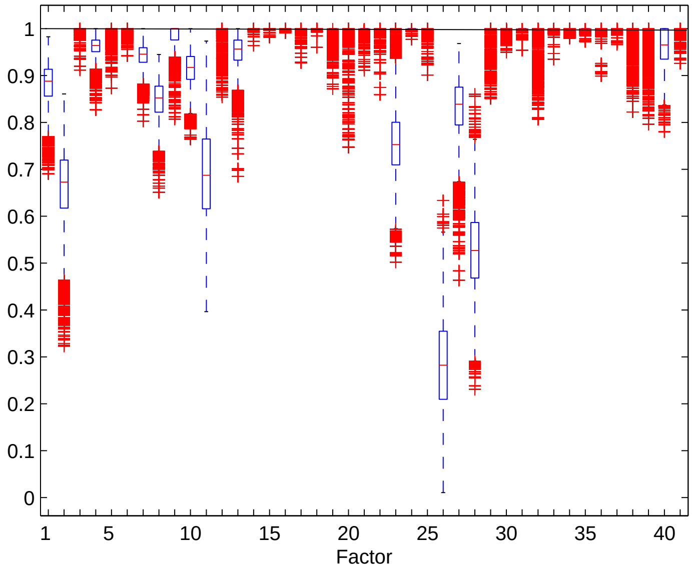
The complicated nature of this code makes it an ideal candidate for the flexible GP model. Schonlau and Welch (2005) use the Wonderland experiment to illustrate an analysis of variance based approach to screening in computer experiments. In their approach, they fit the GP model via a sequential one-at-a-time likelihood approach. The GP predictor is then decomposed into main effects, two-factor interactions, and higher order interactions. The importance of each effect is assessed by looking at its percentage of total functional variance. This parallels other variance-based methods for sensitivity analysis, such as the Sobol' decomposition (Sobol', 1990). Saltelli, Chan and Scott (2000) present a nice overview of this and other sensitivity analysis techniques.

We applied RDVS to identify active factors for this experiment. To implement our approach, we transform the 500 × 41 design matrix to the unit hypercube and standardize the response (HDI). It is suggested by Schonlau and Welch (2005) that no extra error term is required for fitting this code. Following their suggestion, we model the response simply as

$$y(\mathbf{X}) = z(\mathbf{X}),$$

where $z(\mathbf{X})$ is a Gaussian process with the spatial correlation described earlier. Figure 6 shows the posterior distribution of ρ for the 41 factors, with the 10^{th} percentile of the reference distribution drawn. By choosing as active those factors with a posterior median of ρ below the cut-off, we identify 13 of the 41 factors to have an important impact on the response (factors 1, 2, 4, 7, 8, 10, 11, 13, 23, 26, 27, 28, and 40). Schonlau and Welch (2005) identified eight active factors. In our analysis, we find the same 8 factors plus an additional five. At the screening stage, we feel comfortable being more liberal on the identification of active factors, particularly in this setting where the extra factors identified appear to be explainable. For example, Schonlau and Welch (2005) identified sustainable pollution in the southern hemisphere

Figure 6: Posterior distribution of ρ_k for the Wonderland examples



to be an important factor for the HDI. We found this as well, but also identified sustainable pollution in the northern hemisphere to be active.

The differences in results not only depend on the screening technique used, but also on the estimation procedure for fitting the GP model. As mentioned, Schonlau and Welch (2005) use a likelihood based procedure, where one factor at a time is added to the model. Their approach does not explore the model space as extensively as the one proposed here, and thus one should anticipate some differences. To augment our approach, one could follow up our results by looking at main effect and interaction plots and by also calculating variance components (e.g., Schonlau and Welch, 2005) for the 13 active factors. This allows for further assessment of the importance of these factors. The Bayesian probabilistic sensitivity analysis approach of Oakley and O'Hagan (2004) would be appropriate for the framework we consider.

Remark: Generally, the sensitivity measures (e.g., Saltelli, Chan and Scott, 2000) are used to seek understanding of how a model changes with changes to its inputs. One aspect of this is factor screening. In recent work, Oakley and O'Hagan (2004) demonstrate that many popular sensitivity indices can be estimated using a Bayesian approach, which is particularly beneficial in situations where the code is expensive to run. In many ways, the variable selection method we propose complements these sensitivity analysis techniques. The usual approach to the estimation of variance decomposition indices (e.g. Schonlau and Welch, 2005; Oakley and O'Hagan, 2004) requires computing an index for each main effect and two-factor interaction. For example, in the case of the Wonderland model this means evaluation high-dimensional integrals for 41 main effects and 820 two-factor interactions. The advantage of this approach, however, is that it results in effect estimates that can be easily visualized and interpreted. On the other hand, sensitivity indices that attempt to reduce computation by amalgamating the effects of each factor, such as the total effect index (Saltelli, Chan and Scott, 2000; Oakley and O'Hagan, 2004), lose interpretability and only suggest relative importance of inputs, since it is unclear what are "large" or "small" values of this index.

The proposed methodology serves as a cohesive answer to these challenges. It can be used to screen directly or as a precursor to the visualization and ANOVA approaches through identification of factors that are potentially active (in any capacity) with relative ease. Variance components can then be calculated, post-hoc, for main effects and two-factor interactions involving only these active factors (rather than all the factors) so the nature of their effects can be visualized or quantified. Alternatively, the RDVS technique can be used for assessing significance of total effect indices. In the methodology described, we look at the posterior distribution of ρ for each input to assess their importance relative to the posterior distribution of ρ for a dummy factor. There is no particular reason why one could not create a reference distribution of the total effect index instead. We look at the distribution of ρ for simplicity, since it falls directly out of the estimation procedure and captures similar information as the total effect index.

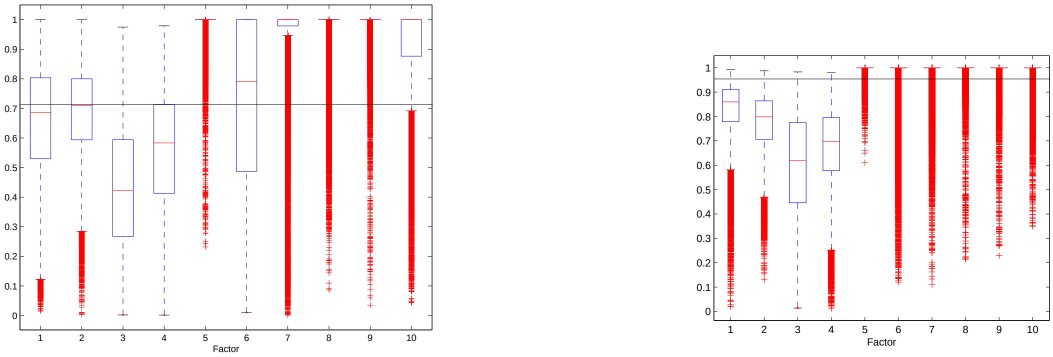
6 Sensitivity to Choice of Prior Distributions

To further understand the performance of RDVS, it would be beneficial to consider its robustness to the choice of prior distributions. Recall from Section 3 that priors are assigned for the GP parameters λ_z, λ_ϵ, and ρ. The prior assigned to λ_z was a gamma distribution with parameters a_z and b_z chosen so that $E(\lambda_z) = 1$. This selection was made to reflect the prior belief that the GP $z(X)$ should account for essentially all of the variability in the standardized response. This is expected in this setting, so we do not explore alternative priors on λ_z.

A gamma prior was also used for the white noise variability λ_ϵ, governed by parameters a_ϵ and b_ϵ. The prior on λ_ϵ specifies the amount of variability in the response that can be attributed to random error. We chose a_ϵ and b_ϵ so that $E(\lambda_\epsilon) = 100$; that is, so that it is expected only about 10% of the response standard deviation is explainable by random error. We also had the additional constraint that $\lambda_\epsilon < 5$, which prevented the white noise component from absorbing any more than about 45% of the response standard deviation at any realization of the MCMC analysis.

To investigate the robustness of RDVS to the choice of prior on λ_ϵ, we try varying the choice of b_ϵ. For fixed a_ϵ, changing b_ϵ allows for adjustments to the mean of this prior distribution. Consider again the linear response function given by (5) in Example 1 of the previous section. The simulation study on this response function is repeated with two alternative prior choices for λ_ϵ. First, a $\Gamma(a_\epsilon = 2.5, b_\epsilon = .0025)I_{[\lambda_\epsilon > 5]}$ prior is used. Under this prior, $E(\lambda_\epsilon) = 1000$, which implies only about 3% of the response standard deviation is expected to be attributable to noise. The same lower bound constraint is kept. An example of the impact on the analysis due to making this change on b_ϵ can be seen in boxplots of the ρ_k posterior distributions given in Figure 7(a). For this plot, the same linear response used for the detailed illustration of RDVS in Example 1 is used. This prior encourages the GP to account for more of the variability in the response, which manifests itself as an increased signal, or more values far from one in the boxplots. However, the added inert variable is given the same prior, and it self-calibrates itself to behave like the other inert factors. As a result, the 10^{th} percentile cut-off of the reference distribution is also farther from one, and the correct variable selections are still made. The results over 1000 simulations (with the response generated by (5)) are given in Table 5. This table shows a slight decrease in the frequency of the correct detection of the first four factors compared to Table 1 of Example 1.

Alternatively, we consider changing the prior on λ_ϵ to encourage more of the variability to be absorbed by the random error component. To do this, a $\Gamma(a_\epsilon = 2.5, b_\epsilon = 0.1)I_{[\lambda_\epsilon > 5]}$ prior on λ_ϵ is used. Given this value of b_ϵ, $E(\lambda_\epsilon) = 25$, so about 20% of the response standard deviation is expected to be in the error. This has the opposite impact on the posterior distribution of the ρ_k as the previous change. In this case, the boxplots corresponding to inert factors are pushed



(a) $b_\epsilon = 0.0025$

(b) $b_\epsilon = 0.1$

Figure 7: Posterior distributions of the experimental variables corresponding to changes in the prior on λ_ϵ.

Table 5: Proportion of times each factor is identified as important in 1000 generations of the response when the prior on λ_ϵ has $b_\epsilon = 0.0025$.

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.568	0.578	0.680	0.566	0.043	0.041	0.028	0.079	0.050	0.067
10^{th}	0.777	0.809	0.877	0.801	0.078	0.081	0.058	0.135	0.101	0.130
15^{th}	0.902	0.909	0.944	0.915	0.108	0.107	0.092	0.194	0.133	0.180

Table 6: Proportion of times each factor is identified as important in 1000 generations of the response when the prior on λ_ϵ has $b_\epsilon = 0.1$.

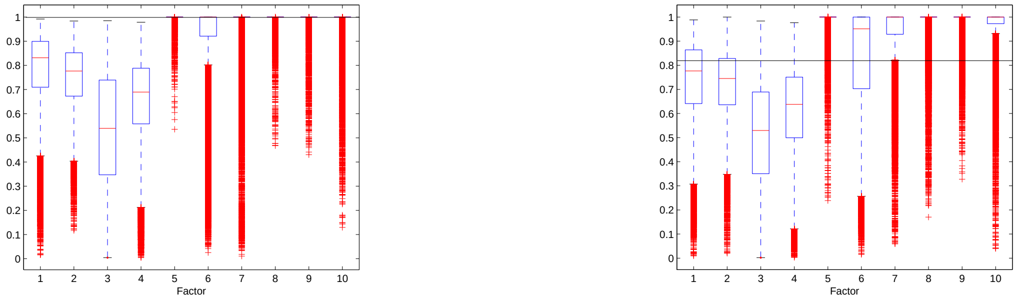
	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.854	0.862	0.895	0.871	0.028	0.027	0.028	0.078	0.040	0.039
10^{th}	0.969	0.979	0.988	0.976	0.043	0.044	0.041	0.107	0.058	0.058
15^{th}	0.995	0.997	0.998	0.995	0.047	0.055	0.046	0.121	0.068	0.063

against one. Mimicking this behavior, the posterior distribution of the added inert factor is also pushed closer to one, as illustrated in Figure 7(b) (again, the same example response was used for this plot). The results from 1000 simulations are displayed in Table 6. In this case, the first four factors are correctly determined to be active with a higher frequency than in Example 1. Overall, changing this prior does have some impact, but due to the self-calibration of the added inert variable, the performance of the RDVS methodology is still quite good.

We next explore the impact of changing the prior for ρ on the methodology. This mixture prior, given in (3), is specified by γ, the prior probability that a factor is active. In all of the previous examples, $\gamma = 1/4$ was taken to be a reasonable value. Consider now two alternative values of γ: $\gamma = 1/10$ and $\gamma = 1/2$. We believe these to be extremities in terms of prior beliefs on effect sparsity. Returning to the linear function given in (5), the simulations are repeated with these varying priors. Again, because the added factor has the same prior information as the other factors, its corresponding posterior distribution still mimics those of the other inert factors in the analysis. Figure 8 demonstrates this point for the same illustrative response used throughout. As can be seen in Tables 7 and 8, the performance of RDVS is quite robust to the prior choice of γ.

7 Cylinder Deformation Application Re-Visited

The examples explored in Section 5 demonstrate that RDVS performs well as a variable selection methodology. In this section, we use it for screening in the Taylor cylinder example in Section 2.



(a) $\gamma = 0.1$

(b) $\gamma = 0.5$

Figure 8: Posterior distributions of the experimental variables corresponding to changes in the prior on ρ_k.

Table 7: Proportion of times each factor is identified as important in 1000 generations of the response when the prior on ρ has $\gamma = 0.1$.

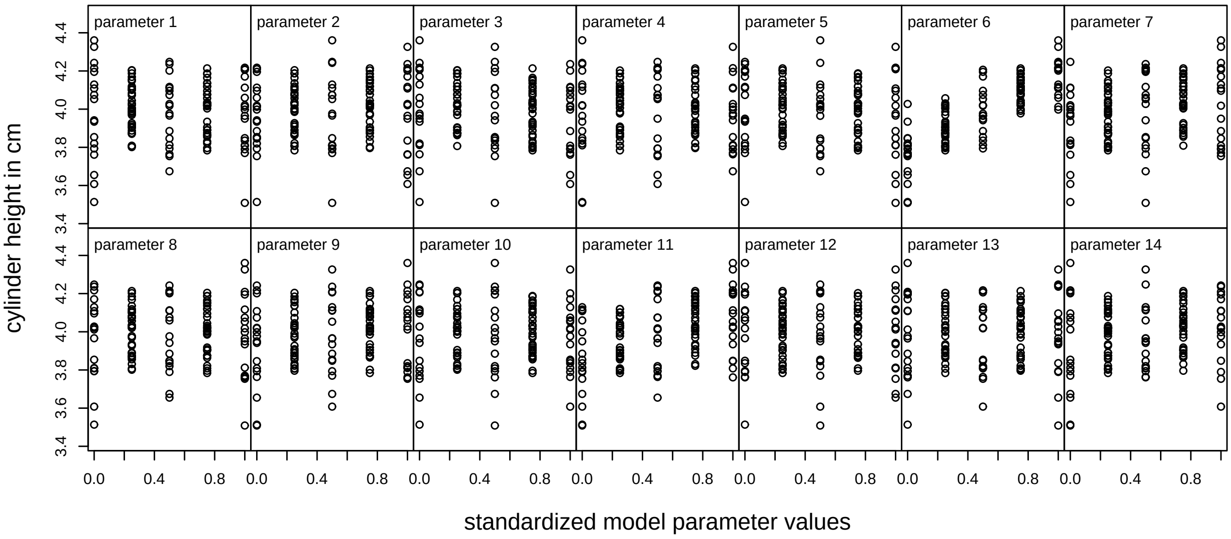
	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.652	0.637	0.742	0.674	0.036	0.019	0.023	0.069	0.030	0.041
10^{th}	0.887	0.878	0.907	0.892	0.054	0.038	0.037	0.096	0.052	0.076
15^{th}	0.963	0.955	0.981	0.966	0.064	0.048	0.047	0.109	0.063	0.093

Table 8: Proportion of times each factor is identified as important in 1000 generations of the response when the prior on ρ has $\gamma = 0.5$.

	Factor									
Percentile	1	2	3	4	5	6	7	8	9	10
5^{th}	0.659	0.675	0.759	0.661	0.037	0.035	0.035	0.114	0.043	0.072
10^{th}	0.844	0.875	0.921	0.868	0.068	0.070	0.074	0.172	0.093	0.120
15^{th}	0.939	0.945	0.974	0.948	0.099	0.110	0.113	0.235	0.148	0.176

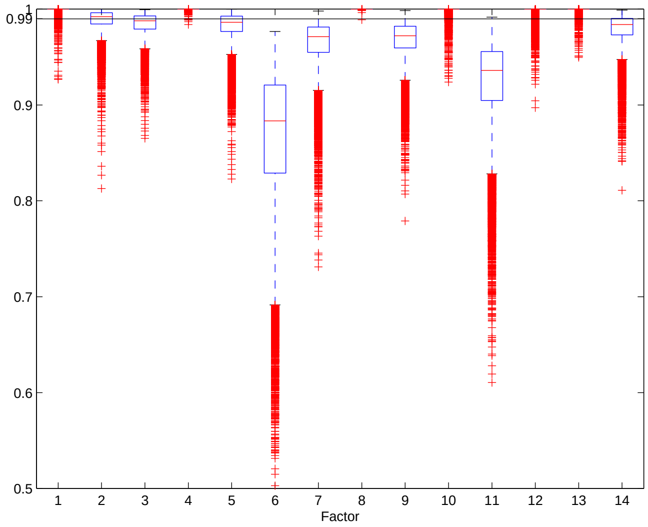
Recall that this model is governed by 14 parameters. A computer experiment was performed based on a five-level, nearly orthogonal array design (Wang and Wu, 1992), which prescribed 118 different input settings at which to carry out the simulation trials. The output from the code is multivariate, and thus there are several measures of deformation one could consider using in the analysis. Since the method presented is for univariate responses, we chose to use the length of the cylinder after impact as our response. Figure 9 shows plots of the simulated cylinder length against the five standardized settings for each of the 14 input factors. From this rudimentary figure it appears that factor 6, which controls how temperature and density affect the plastic stress of the metal, is most important. It is difficult to otherwise distinguish between the factors, so RDVS is used to determine which of the factors are influencing the cylinder lengths after impact. To implement RDVS in this setting, the 118×14 design matrix, X, is (repeatedly) augmented with an additional column and the model outlined in Section 4 is fit. The idea is that this would be the analysis followed if this were a 15 factor experiment. In this case, however, it is known for certain that the 15^{th} factor is inert.

Figure 9: Plots of 118 simulated cylinder heights versus standardized input parameter settings for each of the 14 parameters governing the plastic-elastic flow model used to model the cylinder deformation.



To be comparable, the added factor should be treated in the same manner as the experiment factors in both the design and analysis stages. Thus the added factor should have five level settings, with 23 or 24 trials per setting, corresponding to the original 5-level, nearly orthogonal design matrix. To create the random added column, we begin with a vector which has five equally spaced level settings, $(0, 0.25, 0.50, 0.75, 1.00)$, with 23 copies of each level (i.e., a 115×1 vector). Next, three additional trials from the five level settings are randomly chosen, giving 118 trials for the added factor. The vector is then randomly permuted, resulting in the added column. This procedure is repeated for each of the $m = 100$ added columns.

Figure 10: Posterior distributions of ρ_k for the experiment factors in the Taylor cylinder experiment.



A quick glance at Figure 10 reveals that our initial intuition is confirmed (i.e., factor 6, the impact of temperature and density on the stress rate, is an important factor). When the 10^{th} percentile of the posterior distribution of the median correlation parameter for the inert column is drawn, seven factors are identified as active: factors 3, 5, 6, 7, 9, 11, and 14. Notice that factor 2 is deemed inert since the the posterior median of ρ_2 (0.9921) is larger than the cut-off (0.9909) computed from the posterior distribution of the median correlation parameter for the added factor. It is likely, however, that an experimenter may consider carrying factor 2 forward to the next stage of investigation if the cost of doing so is not prohibitive.

Remark: By using RDVS, the number of experimental factors to be carried on to the next experiment is reduced by half. When conducting subsequent trials, the experimenter will be faced with setting the levels of both the active and inert factors to run the code. The next stage design should give priority to the active factors so the experiment goals (e.g., optimization or response surface estimation) can be met. For an inert factor, one may elect to adjust these as well, thereby allowing for a re-evaluation of the the original screening decisions. Frequently, the cost of adjusting the levels of an input factor cannot be ignored. For instance, if the factor of interest is the mesh threshold in a finite element analysis and is inert over the range explored, one would choose the most coarse mesh (in the range of meshes explored). Changing the level of this factor would mean investigating a finer mesh which, in turn, can cause the simulator to run substantially longer in the future investigation.

8 Concluding Remarks

We have proposed a new method of variable selection for Bayesian Gaussian process models. The basic idea arises from a thought experiment: what would the posterior distribution corresponding to the correlation parameter of an inert variable look like given the data? This question is addressed by including an inert variable in the analysis to find a reference distribution against which to assess the importance of the true experiment variables. This provides an interpretable way to make screening decisions. The simulated examples in Section 5 and real computer experiment analyzed in Section 7 demonstrate the promise of this new approach.

There are a few other issues worth noting. Firstly, it is natural to wonder whether some of the computational effort can be saved by using several added factors per iteration of the methodology, and conducting fewer iteration. After some experimentation, we have found that using more than one added factor in the setting considered here can work quite well. So, if one adds, for example, 4 extra factors, but with only a quarter of the RDVS steps, similar decisions are made. We also found, however, that if one adds too many variables (say 100) one quickly loses power.

Secondly, it is possible to imagine a scenario where ρ_{p+1} converges immediately to 1 for all added columns. As a result, there would be no noise in the posterior distribution of a null median correlation parameter and, therefore, all factors with $\tilde{\rho}_k < 1$ could be deemed active. This can occur when the response is dominated by a few factors, but the factor sparsity assumption fails to hold. That is, when all factors are active to some degree, but only a few of them explain most of the variability. In this case, one is not interested in which factors are inert, but rather which are negligible. In such screening applications, deciding which factors are inert or active may be too liberal a criterion for determining which factors to investigate in further experimentation. To address this, one could *spike* the response with some very small effect based on the added column. That is, the added factor is not inert, but has a small enough impact to be considered negligible (say less that 5% of the unstandardized standard deviation of the response). The posterior distribution of the median ρ_{p+1} would then represent that of a factor which is negligible rather than entirely inert.

Lastly, for future exploration, it would be interesting to see how this method would perform for other models. There are some features of RDVS that point to its success in the particular case of the Gaussian process model that may need to be reconsidered in order to extend the methodology to other models. For example, the number of active and inert factors at each iteration of the MCMC varies because of the draw of ρ_k from the mixture distribution in (3). This has the effect of changing the model size through each iteration. In a regression context, this would amount to entertaining sub-sets of factors rather than the saturated model. In order for the posterior distribution of the white noise component in a regression problem to be unaffected by this procedure, one might have to keep the expected model size constant rather than the factor inclusion probability, γ, constant (as we elected to do). In the Gaussian process case, this was not important because the model used had very little variability associated with the random error, and this error was only included to adjust for numerical jitter and roughness in the response.

References

Box, G.E.P and Meyer, R.D. (1986). "An Analysis for Unreplicated Fractional Factorials," *Technometrics*, **28**, 11-18.

Chipman, H.A., George, E.I. and McCulloch, R.E. (2001). "The Practical Implementation of Bayesian Model Selection," *IMS Lecture Notes - Monograph Series*, **38**.

Clyde, M. (1999). "Bayesian Model Averaging and Model Search Strategies (with discussion)," *Bayesian Statistics 6*, Bernardo, J.M. , Dawid, A.P. , Berger, J.O. and Smith, A.F.M. Eds. Oxford University Press, 157-185.

Kennedy, M. and O'Hagan, A. (2001). "Bayesian Calibration of Computer Codels (with discussion)," *Journal of the Royal Statistical Society, Series B*, **63**, 425-464.

George, E.I. and McCulloch, R.E. (1993). "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, **88**, 881-889.

Goldstein, M. and Rougier, J.C. (2004). "Probabilistic Formulations for Transferring Inferences from Mathematical Models to Physical Systems," *SIAM Journal on Scientific Computing*, to appear.

Johnson, M.E., Moore, L.M., and Ylvisaker, D. (1990). "Minimax and Maximin Distance Designs," *Journal of Statistical Planning and Inference*, **26**, 131-148.

McKay, M.D., Beckman, R.J. and Conover, W.J. (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Coded," *Technometrics*, **21**, 239-245.

Milik, A., Prskawetz, A., Feichtinger, G. and Sanderson, W.C. (1996). "Slow-fast Dynamics in Wonderland," *Environmental Modeling and Assessment*, **1**, 3-17.

Preston, D.L., Tonks, D.L., and Wallace, D.C. (2003). "Model of Plastic Deformation for Extreme Loading Conditions," *Journal of Applied Physics*, **93**, 211-220.

Oakley, J.E. and O'Hagan, A. (2004). "Probabilistic Sensitivity Analysis of Complex Models: A Bayesian Approach," *Journal of the Royal Statistical Society, Series B*, **66**, 751-769.

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P (1989a). "Design and Analysis of Computer Experiments," *Statistical Science*, **4**, 409-435.

Sacks, J., Schiller, S.B. and Welch, W.J. (1989b). "Designs for Computer Experiments," *Technometrics*, **31**, 41-47.

Saltelli, A., Chan, K. and Scott, E.M. (2000) (eds.) *Sensitivity Analysis.* New York: Wiley.

Schonlau, M. and Welch, W.J. (2005). "Screening the Input Variables to a Computer Model via Analysis of Variance and Visualization," *Screening Methods for Experimentation in Industry, Drug Discovery and Genetics* (Eds A. M. Dean and S. M. Lewis), Springer Verlag.

Sobol', I.M. (1990). "Sensitivity Estimates for Nonlinear Mathematical Models," *Matematicheskoe Modelirovanie*, **2**, 112-118 (translated as: Sobol', I.M., (1993). "Sensitivity Analysis for

Non-linear mathematical Models," *Mathematical Modelling and Computational Experiment*, **1**, 407-414).

Wang, J.C. and Wu, C.F.J. (1992). "Nearly Orthogonal Arrays with Mixed Levels and Small Runs," *Technometrics*, **34**, 450-456.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., Morris, M.D. (1992). "Screening, Predicting, and Computer Experiments," *Technometrics*, **34**, 15-25.

Wu, C.F.J., Hamada, M. (2000). *Experiments; Planning, Analysis, and Parameter Design Optimization*. Wiley, New York.